DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models

Year: 2023 | Citations: 534 | Authors: Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang

---

Abstract

Generative Pre-trained Transformer (GPT) models have exhibited exciting progress in their capabilities, capturing the interest of practitioners and the public alike. Yet, while the literature on the trustworthiness of GPT models remains limited, practitioners have proposed employing capable GPT models for sensitive applications such as healthcare and finance -- where mistakes can be costly. To this end, this work proposes a comprehensive trustworthiness evaluation for large language models with a focus on GPT-4 and GPT-3.5, considering diverse perspectives -- including toxicity, stereotype bias, adversarial robustness, out-of-distribution robustness, robustness on adversarial demonstrations, privacy, machine ethics, and fairness. Based on our evaluations, we discover previously unpublished vulnerabilities to trustworthiness threats. For instance, we find that GPT models can be easily misled to generate toxic and biased outputs and leak private information in both training data and conversation history. We also find that although GPT-4 is usually more trustworthy than GPT-3.5 on standard benchmarks, GPT-4 is more vulnerable given jailbreaking system or user prompts, potentially because GPT-4 follows (misleading) instructions more precisely. Our work illustrates a comprehensive trustworthiness evaluation of GPT models and sheds light on the trustworthiness gaps. Our benchmark is publicly available at https://decodingtrust.github.io/ ; our dataset can be previewed at https://huggingface.co/datasets/AI-Secure/DecodingTrust ; a concise version of this work is at https://openreview.net/pdf?id=kaHpo8OZw2 .